Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Candidate genes and bioinformatic analysis of biological pathways for epistatic regulation of growth in chicken

Muhammad Ahsan

Department of Animal Breeding and Genetics
Examensarbete 332
Uppsala 2010

Master's Thesis, 30 HEC

One-Year Master's Programme in Biology – Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Candidate genes and bioinformatic analysis of biological pathways for epistatic regulation of growth in chicken

Muhammad Ahsan

Supervisors:
Stefan Marklund, SLU, Department of Animal Breeding and Genetics
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Examiner:
Örjan Carlborg, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Name of series: Examensarbete 332
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: complex traits, QTL mapping, SNP, mutation, divergent selection, KEGG, UCSC, Ensembl, Galaxy, GO, conservation

CONTENTS

ABSTRACT

Selective breeding programmes for desired phenotypes in animals provide a good resource for researchers to understand the genetic basis underlying those phenotypes. Most traits of economic importance in animals are quantitative traits, which are controlled by genetic as well as environmental factors. Previous studies have shown that the mapping of the genomic regions containing genetic factors, which control the diverse phenotypes in complex traits of interest, resulted in the successful identification of causative mutations. The same forward genetics approach, which endeavours to identify the genetic basis of phenotypic diversity, was followed in this study. To identify and functionally evaluate the mutations controlling diverse phenotypic differences between the High-weight and Low-weight chicken lines, the previously fine mapped interacting quantitative trait loci (QTL) for growth in chicken were scanned for the presence of genes. Gene databases (Ensembl, RefSeq) and gene prediction tools (Genscan and N-Scan) were used to reveal as many genes as possible. To identify candidate genes for growth, the conservation pattern of those genes was studied using the phastCons programme in the PHAST package. The genes most conserved in evolution were identified and selected for further analysis. The evidence based, well-supported and annotated Ensemble conserved genes were identified as candidate genes for growth in general. Gene ontology terms for the candidate genes were obtained from the Gene Ontology (GO) database to study the molecular and biological functions of those candidate genes. KEGG biological pathways were mapped for the presence of identified candidate genes. The candidate genes were found having roles in many biological pathways. The biological pathways that contained candidate genes from multiple interacting QTL were identified as candidate pathways. Major candidate pathways include the Mitogen Activated Protein Kinase (MAPK) signalling pathway, which affects growth in general, and Adipocytokine signalling pathway, which imparts its anorectic effect through the leptin hormone. Further studies are required to analyse the functional effects of those candidate genes and pathways and eventually to identify the mutations affecting growth.

INTRODUCTION

Animal breeding for traits of interest has a proven potential to achieve large phenotypic effects in a very short time frame from an evolutionary perspective. This does not only offer great possibilities to transform the animal phenotype in the direction desired but also provide means for biological research to understand the genetics behind those traits. Most of the traits of economic importance in animals are complex in nature. Those traits are often controlled by genetic factors as well as by environmental factors. Genetic factors involve variation in one or more genes causing variation in phenotype whereas environmental factors include other factors with effect on the phenotype. Genes impart their additive effects on the phenotype but do not explain all the phenotypic variation. Inter-locus interactions, known as epistasis, play a vital part in explaining the phenotypic variation in many complex traits.

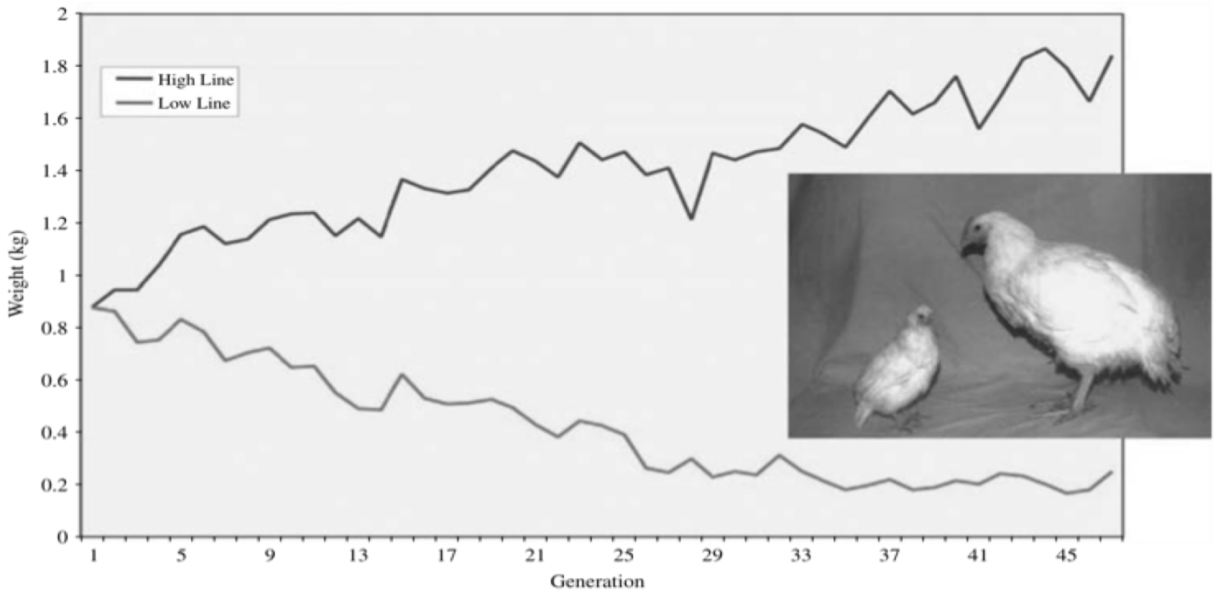
Chicken is a vertebrate animal model for the study of human biology and medicine and is the first avian species to have its genome sequenced. It is a main source of human food and protein. Chicken is also useful for genetic mapping and QTL analysis due to its relatively short generation interval and high fertility, which means that it is possible to relatively rapidly breed several generations of large families.

Chickens have typically been under selection for complex trait of growth to increase meat production and/or for egg production and quality. The commercial populations provide a resource also for genetic studies. However, to study the selection process itself, the phenotypic effects and biological differences between selected populations might also be the main purpose of a breeding program. In 1957, Paul Siegel in USA started a breeding program with bi-directional selection in chicken for body weight at 56 days of age (Dunnington and Siegel, 1996). The base chicken population was developed from crosses between seven partially inbred lines of White Plymouth Rock. The selection during more than 40 generations has resulted in two divergent chicken lines namely the High-Weight (High-Line) and the Low-Weight (Low-Line). The average body weight at 56 days of the base population was 793 grams (Figure 1). In contrast, the average weights of the High-Line and Low-Line after 40 generations were 1412 grams and 170 grams, respectively, at 56 days (Carlborg et al, 2006). Several differences in physiological parameters and metabolic traits between these two lines can contribute to the observed differences in body weights. The phenotypic differences are overwhelming in terms of difference in growth rate and other metabolic traits like appetite, fat deposition and immune response between those two lines. The difference in feed intake is remarkable where the High-Line has become hyper-phagic and feed restrictions are applied to control their body weight after selection and on the other hand, the Low-Line shows very low appetite and takes feed almost inadequate for survival (Jacobsson et al, 2005). The High-Line is found to have lower immune response than the Low-Line and it is presumed that it may be due to the competition for resources between growth and immune system (Cheema et al, 2003).

Chicken's genomic DNA is organized into 39 pairs of chromosome (38 pairs of autosomes and 1 pair of sex chromosome) inside a nucleus. Chicken is the first avian whose genome has been sequenced. In Ensembl database the size of chicken genome is 1,050,947,331 base pairs, which is approximately one third of the human genome (3,274,571,503 base pairs). Ensembl also reveals

17,934 genes (14,959 Known protein-coding) of all kinds (known and novel protein-coding genes, Pseudogenes, RNA genes) in chicken genome as compared to 51,737 genes of all kinds (21,257 Known protein-coding) in human genome (Ensembl 59).

Figure 1: Body weights of male chickens of High- and Low Lines at the age of 56 days. Also shown chickens from both lines in 37th generation after selection (Jacobsson et al. Genetical Research (2005). (Reproduced with permission)



Multiple experimental strategies are devised to identify genomic regions containing genes that cause phenotypic variation in complex traits, which include Genome Wide Association Studies (GWAS) and Quantitative Trait Loci (QTL) mapping. Mapped genomic regions contain at least one mutation that has a functional role in the trait of interest.

Quantitative trait loci (QTL) can be defined as regions in DNA that harbour one or more genes controlling complex traits of interest. Identification of those regions that are causing phenotypic variation in the complex trait of interest is the aim of QTL mapping. The total genetic variation in a population is the sum of the contribution from individual QTL and their interactions. Series of experimental crosses (mainly Backcross and F2 Intercross) in a mapping population may reveal the underlying genes controlling the trait of interest (Broman, 2001). The detection of co-segregation of alleles of genetic markers and trait phenotypes results in the detection of QTL in those crosses (Carlborg, 2002). A genetic marker is a gene or region of DNA with known sequence polymorphism whose genotype can be determined by using a suitable method of choice whereas a genetic map shows all those known markers at their relative positions on chromosomes, with markers on the same chromosome separated by certain genetic distances (centiMorgan) based on the observed recombination rate between the markers in a family material. To maximise the chances to detect co-segregation between genetic markers and QTL all chromosomes and the mitochondrial genome should be well covered with informative (segregating) markers and the phenotypes of interest need to be segregating within the pedigree material / experimental cross. In crosses between phenotypically divergent inbred lines (lines highly homozygous at marker loci and QTL) it is more likely that segregation of QTL may be detected (Carlborg, 2002). Table 1

shows that the backcross progeny of phenotypically divergent inbred parental lines may have one of the two possible marker genotypes which makes it the simplest method whereas in case of F2 intercross there are more than two marker genotypes in the progeny allowing the detection of QTL of additive, dominance, and epistatic effect.

Table 1. Backcross and F2 Intercross experiments showing marker genotypes

Backcross	Genotype	F2 Intercross	Genotype
Parent	AA X BB	Parent	AA X BB
F1	AB	F1	AB
Backcross (AA X AB)	AA or AB	F2 (AB X AB)	AA or AB or BB
Backcross (BB X AB)	BB or AB		

There are different statistical methods for mapping QTL and Analysis of variance (ANOVA) at the marker loci (marker regression) is the simplest (Broman, 2001). In ANOVA, for each marker the backcross progeny is divided into two groups with respect to their genotypes at the marker loci, followed by the comparisons of the phenotypic distributions between the two groups. Distributions that are having phenotypic values approximately the same indicate that a marker is not linked to a QTL. A marker linked to a QTL is indicated by the significant differences in the phenotypic distribution of the two groups. For comparison of distributions, t-statistic may be calculated in a backcross experiment where progeny may have only one of the two marker genotypes and F-statistic may be calculated in the intercross experiment where there are more than two marker genotypes possible. In this way the genome is searched for finding significant associations between markers and the phenotypes.

Currently the most popular technique in QTL mapping is Interval mapping (with some modifications to the original method of Lander and Botstein 1989), which estimates marker genotypes and association in the interval at each position between back-to-back marker pairs. Likelihood ratios or the Logarithms of the Odds (LOD) score are calculated as a statistical support for QTL in interval mapping.

Genetic effects of the QTL can be modelled in different ways. The most common single locus models are for additive effects and dominance effects. The additive model measures an allele substitution effect by changing a low effect allele for a high effect allele where the heterozygote shows an intermediate phenotype to both homozygote phenotypes. In the dominance model the heterozygote phenotype is biased towards one of the homozygote phenotypes. Once QTL mapped it becomes possible to detect and measure effects of interactions between genotypes at different loci (two or more) known as epistasis.

For positional cloning and identification of genes, which underlie the complex traits, fine mapped quantitative trait loci (QTL) data serves as a genetic resource of foremost importance for the researchers (Liu et al, 2007). Identification of QTL is critical to find and explain the biochemical mechanism of complex traits (Broman, 2001). Jacobsson et al. (2005) found that the difference in growth between the High- and Low-Lines of chicken is regulated by 13 different QTL on 9 different chromosomes and all showed minor additive effects except two growth QTL (G11 on chromosome 20 and G13 on chromosome 28) which showed negative over-dominance resulting in reduced growth in heterozygotes. However, interactions between several of these QTL explain a larger proportion of the phenotypic differences (Carlborg et al, 2006). Recent fine mapping have

confirmed the majority of those previously found QTL and also the interactions between the QTL on chromosome 3, 4 and 7 (Besnier et al, in preparation). So it is not only the additive effects of the gene(s) in each QTL that control the QTL effects on phenotypes but also the epistatic interactions between QTL on different chromosomes.
In the current study we concentrated on the three epistatic QTL on chromosome 3, 4 and 7, which have a major effect on growth (Carlborg et al, 2006, Besnier et al. (in preparation)), in order to trace the genetic basis (genes and mutations) of the complex phenotype of growth

Most genome projects require genome sequences to start with functional genomics research. Today the complete reference genome sequences of several organisms including chicken are available. This provides us the opportunity to identify almost all the genes in those genomes. To identify genetic polymorphism affecting traits of interest certain populations or individuals are being re-sequenced. In our case the High- and the Low-Line chickens were re-sequenced to reveal the genetic polymorphism that could be contributing towards phenotypic variation between those phenotypically divergent lines. To sequence or re-sequence the genomes, the next generation sequencing (NGS) technologies are providing high-throughput methodologies in an affordable way for most functional genomic projects and that allows researchers to detect mutations and polymorphism that exist in the interesting genome regions in the populations that are under study. These platforms produce tens of thousands of megabases of sequence data per run. The technologies are currently being provided mainly by 454 (Roche), Solexa (Illumina) and SOLiD (ABI). These NGS technologies have gained preference over the state-of-the-art Sanger sequencing technology few years ago although compromising read length.
Sanger sequencing uses chain termination method. The DNA fragments are cloned *in vivo*, usually inside a bacterial host. To synthesize complementary DNA strands the DNA polymerase, DNA primer and deoxynucleotides (dATP, dGTP, dCTP and dTTP) are added to DNA clones. The dideoxynucleotides (ddATP, ddGTP, ddCTP and ddTTP) are also added, which serve as chain terminators. The polymerase incorporates the complementary deoxynucleotides to a growing chain of complementary oligonucleotide and a dideoxynucleotides incorporated at random abruptly terminates the synthesis of the oligonucleotide chain. In this way several complementary strands of different lengths of a DNA template are synthesized having detectable dideoxynucleotides at their 3′ ends. This reaction generates all possible complementary oligonucleotides varying in size by one nucleotide but starting with the common 5′ base of a template DNA strand. The synthesized DNA strands of varying lengths are size-separated using gel electrophoresis. From the ladder of bands formed due to electrophoresis the 3′-end dideoxynucleotide is recognized for the template DNA strand. The sequence is read from the bands as shown in Table 2.

Table 2. Sanger Sequencing using chain termination methodology

Synthesized complementary oligonucleotides 5′----------->3′	Dideoxynucleotide at the 3′ end	Complementary sequence
ATCGTCC	C	ATCGTCC
ATCGTC	C	
ATCGT	T	
ATCG	G	
ATC	C	
AT	T	
A	A	

Basic principle of chain termination in the Sanger sequencing remained the same over time but technological advances in the original methods proved to decrease errors in sequencing along with longer reads. Current methods on ABI 3730xl Sanger sequencing instrument can produce up to 96 Kb of sequenced data in a 3-hour run in the form of reads of size ≥ 900 bp.

The next generation sequencing technologies offer affordable sequencing alternatives to the costly Sanger sequencing, which are also high-throughput, producing billions of bases of sequence data in a single run. The 454 / Roche technology was the first of the NGS technologies to be marketed and it is a massively parallel pyrosequencing scheme of sequencing-by-synthesis technique (Margulies et al, 2005). The DNA amplification is *in vitro* in contrast to the Sanger's *in vivo*. For DNA amplification Emulsion PCR is used in which each DNA fragment attached to a bead through an adapter is enclosed in an emulsion droplet. Each droplet containing a different template serves as cloning reactor for producing several thousands of clones per template attached to the bead. A subsequent pyrosequencing reaction in very minute wells for every bead allows sequencing of the clones. The reaction is associated with the release of inorganic pyrophosphate (PPi) due to the incorporation of a complementary nucleotide in the oligonucleotide chain (being synthesized) and the resultant light emission by an enzyme is measured to generate a "Pyrogram". Four deoxynucleotides are added one at a time in a certain order repeated in several hundred cycles in the reaction to the immobilized template DNA and whenever it is incorporated, light emitted by enzyme is measured. A Pyrogram shows the correct order of the nucleotide incorporation and hence the sequence of the template strand. Currently the 454 can produce up to 80-120 Mb of sequence data in the form of 200-300 bp longer reads in a run of 4-hour. Millions of reads produced are then aligned with the reference genome or assembled into a new genome sequence *de novo* without needing any reference genome to align with. Biological applications for 454 are *de novo* assemblies of bacterial and insect genome. Long reads in 454 improve mapping in repetitive regions but error rates are also higher.

The Illumina (Solexa) technology relies on cloning-free DNA amplification, which is done by attaching single-stranded DNA fragments to a transparent surface known as flow cell. Single-stranded DNA fragments (200 bp) are attached to the surface using 5′ and 3′ adapters. Those adapters are complementary to the primers on the surface. These fragments are sequenced one base at a time using fluorescently labeled dideoxynucleotides, which after incorporating and being read are able to allow incorporation of another nucleotide in the next cycle. The four nucleotides are added in a certain order in 50 or more cycles, starting with a primer complementary to one of

the adapter in first cycle and then starting with the primer to the other adapter in the second cycle to sequence a mate-pair read (50-mers sequenced on both sides of the DNA template). The DNA templates are sequenced in a massively parallel fashion on the surface using a DNA sequencing-by-synthesis approach. The 1G-genome analyzer from Illumina, inc., can generate 35-bp single reads and produce at least 1 Gb of sequence data per run in 2-3 days. The biological applications include variant discovery by whole-genome re-sequencing and whole-exome capture. Illumina is currently the most widely used technology in NGS.

ABI/SOLiD enables massively parallel sequencing by hybridization-ligation of amplified clones attached to beads. Emulsion PCR of single-stranded DNA is carried out to construct sequencing libraries which is similar to that used in the 454 technique. The DNA clones on the glass surface are sequenced using 16 dinucleotide combinations labeled by four different fluorescent dyes (each dye representing 4 dinucleotides) by sequential rounds of hybridization and ligation. The sequence is read every fifth base at a time and the nucleotide sequence is determined from the colour analysis resulting from the two successive ligation steps. The SOLiD instrument can produce 1-3 Gb of sequence data per an 8 day run in form of 35 bp reads. The biological applications include variant discovery by whole-genome re-sequencing and whole-exome capture. The two-base encoding scheme in SOLiD provides effective error correction mechanism

For genome re-sequencing of the High-Line and the Low-Line we used the SOLiD system. This technology uses two-base encoding scheme, which makes the distinction between a sequence polymorphism and a sequencing error albeit short sequence reads. SOLiD sequencing produces maximum data per run of all next generation techniques, 1-3 Gb of sequence data in form of 35 bp long sequence reads. The 454-pyrosequencing approach incorrectly estimates the length of indels whereas Illumina is better and effective than 454 in sequencing indels and produces 1 Gb of sequence data per run. On the other hand, smaller read size of 30-40 bp in Illumina results in unresolved short sequence repeats (Morozova & Marra, 2008; Metzker, 2010).

This study is a part of a project with an overall aim to reveal and explain genetic architecture that contributes to the phenotypic differences between the High-Line and the Low-Line. The current paper is an interim report of our investigation of the presence of functional elements and conserved sequences and genetic variation within three epistatic QTL regulating growth in chicken.

MATERIALS AND METHODS

The Virginia Chicken Lines

The chicken lines bi-directionally selected from the same base population for body weight at 56 days of age, namely, the High-Line and Low- Line have been developed and are being maintained by Paul Siegel at Virginia Polytechnic Institute and State University in Blacksburg, Virginia, USA (Dunnington & Siegel, 1996). For the current study genome re-sequencing data from these lines are available (refer to section Genome re-sequencing data). These High- and Low-Line in generation 41 served as parental population to produce F2 generation through a design of reciprocal intercrosses for QTL analysis of growth (Jacobsson et al, 2005)

QTL analysis of growth

For QTL analysis of growth, Jacobsson et al. (2005) used reciprocal intercrosses between the chickens in generation 41 of the High-Line and the Low-Line to produce 874 F2 chickens. 145 microsatellite markers were grouped into 25 linkage groups to form a genetic map (Jacobsson et al, 2004). A regression based analysis (Haley et al, 1994) was then performed for QTL analysis to detect 13 QTL that are affecting growth and except Growth11 and Growth13, all QTL showed mainly additive effects.

Epistatic QTL analysis of growth

Carlborg et al. (2006) used the same F2 intercross between the High-Line and Low-Line chickens for epistatic QTL mapping. The same genetic markers and linkage groups were used to detect epistasis using a statistical model which contained additive, dominance and all pair-wise epistatic effects of QTL pairs and the fixed effect of sex. Besnier et al. (in preparation) further confirmed the QTL on chromosome 3, 4 and 7, which include epistatic interactions with major influence on the regulation of growth.

Genome re-sequencing data

Genome re-sequencing data with 5X average depth coverage in each of the High-Line and the Low-Line were obtained as follows: One pool of genomic DNA from each line, with seven males and four females represented in each pool, was re-sequenced using *Applied Biosystems* SOLiD v2 (next-generation sequencing technology) according to the manufacturer's protocols. The re-sequencing was performed using a genomic fragment library and 35 bases per read. The reads were aligned to the Red Jungle Fowl's reference genome assembly (version 2.1, Washington University) using the MAPREADS program by *Applied Biosystems (Life Technologies)*, which maps SOLiD system sequencing reads to a reference genome sequence (Rubin et al, 2010). *AB Resequencing Analysis Pipeline (Corona Lite)* is data analysis software also from *Applied Biosystems* that analyses and aligns SOLiD reads against reference genome sequences and also calls SNP. Three non-reference reads with different start bases were used as a SNP detection threshold in *Corona Lite pipeline*. Thus, the SNP detection process requires that a given non-reference base is read by the software multiple times because the sequencing data might contain some errors.

Gene identification

The epistatic QTL on chromosome 3, 4 and 7 were scanned for known and predicted genes and transcripts. Gene repositories (gene databases and gene prediction tools) of RefSeq, Ensembl, Genscan, and N-Scan were used to extract the genes and transcripts. All these databases and tools for genes and transcripts refer to the chicken's reference genome assembly WUGSC 2.1/galGal3 of May 2006. The UCSC Genome Browser and Tables, Ensembl's BioMart, and Galaxy were used for extracting the genes and gene related information (genomic sequences, coding sequences, translated proteins, protein domains, ontology) from the QTL.

UCSC Genome Bioinformatics

Today there exists a lot of bioinformatics integrated resources freely available on the web to access genomic information but Ensembl and UCSC Genome Browsers are acknowledged as the most popular and comprehensive (Barnes, 2010). All the genome browsers provide fundamental information on genomes like genes, regulatory regions, evolution, variation, homology and structure to list a few.

University of California Santa Cruz (UCSC) Genome Bioinformatics website (http://genome.ucsc.edu/) was developed and is being maintained by University of California's Genome Bioinformatics group (Kent et al, 2002; Karolchick et al, 2004; Rhead et al, 2010). Reference genome sequences of a variety of species including chicken are contained in a centralized and automatically annotated UCSC database. It provides Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) for researchers to visualize the annotated genes on chromosomes by zooming and moving around in specified genome coordinates in a particular genome assembly. Genome Browser provides different levels of annotations from a single nucleotide to a single gene and to a whole chromosome for many species including vertebrates and it addresses a whole range of biological problems in evolution, comparative genetics and bioinformatics (Pevsner, 2009). Custom annotated named tracks can be created for any genomic feature e.g., SNP by specifying chromosome number and genome coordinates. The researchers can upload their own genomic data in custom tracks to visualize and annotate them in the Genome Browser. The underlying genomic information in Genome Browser and custom tracks can be extracted using UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) for further analysis. UCSC Genome Browser also provides comprehensive supporting evidence (e.g., EST evidence, cDNA evidence, Protein evidence) for known and hypothetical genes. For a novel gene a conserved spliced EST is a good supporting evidence (Barnes, 2010).

After identifying all the possible genes in the QTL, logically the next step should be to investigate each gene through literature search, ontology and homology information from sequences of other species for any of their functional roles in the phenotype of interest.

Genes and Gene Prediction in UCSC

The genes and transcripts from Ensembl, RefSeq, Genscan and N-Scan were extracted using the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) by selecting appropriate parameters from the drop down lists. The parameters in our case were as follows:

clade: *Vertebrate*

genome: *Chicken*

assembly: *May 2006 (WUGSC 2.1/galGal3*

group: *Genes and Gene Prediction Tracks*
track: *Ensembl Genes* or *RefSeq Genes* or *N-Scan Genes* or *GENSCAN* (select one at a time)
table: *ensGene* or *refGene* or *nscanGene* or *genscan* (select track relevant table)
region: Select *position* and define the genome region of QTL (e.g., chr3:28,768,446-39,739,740)
output format: Select *Custom Track* for viewing genes in Genome Browser or select *BED-browser extensible data* to send output to *Galaxy*
send output to: Select *Galaxy* if output to be sent to Galaxy web system
file type returned: Select *Plain text*
Press *get output* button with all other parameters set to their default values.
On the next screen select *Whole Gene* and press *Send query to Galaxy* or *get custom track in genome browser* with all default values set on the form.
Table browser outputs a tabular file enlisting all the genes along with their genome positions, exon counts, transcription start and end positions etc. Instead of whole genes researchers can extract separately different regions of genes e.g., exons, introns, UTR, coding exons and gene flanking regions.
Sessions of Genome Browser can be created and saved for future reference. A personal account on the UCSC website can be created to save sessions and custom tracks. The output data of genes from all four tracks were transferred to the collaborating Galaxy web system for conservation analysis.

Most Conserved in Evolution

We also took advantage of the Comparative Genomics' track in the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) to extract the predicted most conserved elements in the QTL based on multiple alignment data from seven vertebrate species including chicken, human, mouse, rat, opossum, zebra fish and X. tropicalus. We were interested in identifying those genes, which contained the genomic regions that probably remained most conserved during evolution. For that reason we retrieved the most conserved elements in the QTL using the UCSC Table Browser and intersected them to identify those genes that contained those most conserved elements. The phastCons tool in PHAST (PHylogenetic Analysis with Space/Time) software package predicts most conserved elements (Siepel et al, 2005). The phylogenetic hidden Markov model, a probabilistic model, predicts conserved segments of sequences from multiple alignment data. Pairwise alignments through the Blastz (Chiaromonte et al, 2002; Schwartz et al, 2003) programme for all of seven species are produced first. From those pairwise alignments a multiple alignment are developed using the Multiz (Blanchette et al, 2004) programme. The phastCons programme is then run on that alignment with the most conserved option to extract the most conserved elements (Siepel et al, 2005). Every conserved element has a log-odds score between 0 and 1000. The score is directly proportional to the conservation probability of the conserved element. In our case we used a middle value of 500 for extracting those elements, which had higher probability of being remained conserved in evolution. That, in turn, helped us to identify those genes (containing the most conserved elements of higher conservation probability), which probably remained most conserved in evolution.
Conservation is considered probably the most important information to elucidate genome functions because sequence conservation shows preserved functions during evolution (Barnes, 2010).

We retrieved the most conserved elements in QTL in UCSC Table Browser by setting the following parameters:
clade: *Vertebrate*
genome: *Chicken*
assembly: *May 2006 (WUGSC 2.1/galGal3*
group: *Comparative Genomics*
track: *Most Conserved*
table: *phastConsElements7way*
region: Select *position* and define the genome region of QTL (e.g., chr3:28,768,446-39,739,740)
filter: Press *create* and on the next screen set the value of *score >=500* and submit
output format: Select *Custom Track* for viewing conserved elements in Genome Browser or select *BED-browser extensible data* to send output to *Galaxy*
send output to: Select *Galaxy* if output to be sent to Galaxy web system
file type returned: Select *Plain text*
Press *get output* button with all other parameters set to their default values
On the next screen press *Send query to Galaxy* or *get custom track in genome browser* with all other parameters set to their default values
Table browser outputs a tabular file enlisting all the conserved elements along with their genome positions and lod score ($\geq$ 500).

Ensembl / BioMart

Ensemble (www.ensembl.org) is one of the leading genome browsers to visualize and retrieve the integrated information of genomes of chordates (Flicek et al, 2010). It provides resources for genome function, evolution and variation along with the gene annotations. Ensembl genes are supported by comprehensive supporting evidence of known ESTs, cDNA, and translated proteins. The Ensembl genomic data are easily accessible through its genome browser for visualization. The BioMart data-mining tool (http://www.ensembl.org/biomart/martview/) is used for retrieving genomic information from the underlying Ensembl genomic databases (Haider et al, 2009). European Molecular Biology Laboratory (EMBL-EBI) and Wellcome Trust Sanger Institute jointly developed Ensembl for creating and maintaining automatic annotation of eukaryotic genomes including chicken. This project is mainly focused on eukaryotic genomes especially vertebrates. Ensembl is widely considered a standard biological database.

In our case we used the BioMart tool to retrieve the Gene Ontology (GO) annotations of candidate Ensembl genes. Genomic regions such as exons, introns, UTR and flanking regions of Ensembl genes for further analysis were also obtained using BioMart.

Galaxy

The UCSC Table Browser and Ensembl BioMart provide flexible and user-friendly interfaces for simple queries to a fairly large genomic data. But to handle complex queries for larger genomic and genetic datasets, Galaxy (http://main.g2.bx.psu.edu/) is considered a very helpful set of integrated tools to do computational analysis without even doing any programming and only requiring an online web browser (Woollard, 2010). Galaxy is hosted by the Center for Comparative Genomics and Bioinformatics in the Huck Institutes of the Life Sciences at Penn

State University (Goecks et al, 2010; Blankenberg et al, 2010). It is a web-based platform for all kinds of genomic research offering analysis tools for Alignments, Evolution, NGS (next generation sequencing) and Metagenomics to list a few. File manipulation tools such as Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group and Operate on Genomic Intervals are some of the tools which are very helpful to take on files without even knowing any programming. It lets you to perform computational analysis such as by defining workflows to repeat the same set of queries on different genetic datasets.. The genomic data from the collaborating genome browsers such as UCSC and BioMart can be imported by Get Data tool. User data can also be uploaded and analysed in Galaxy. The results acquired from the data analysis remain on the online system and can be accessed from anywhere and shared with anyone. The genomic information of genes and most conserved elements in epistatic QTL from UCSC were imported in Galaxy to identify the genes that probably remained most conserved. The imported files were manipulated for conservation analysis with its Text Manipulation tools, Filter and Sort tools, Join, Subtract and Group tools, and Operate on Genomic Intervals. The intersection tool (Operate on Genomic Intervals) identified the genes containing the most conserved elements by detecting overlaps between the gene coordinates and conserved element coordinates.

Ab initio Gene Prediction

After sequencing and assembling the genome of a species the next step would be to annotate the genes in the genome. Protein coding sequences comprising genes can be discovered in the genome taking into account the common features of protein-coding transcripts. This process of gene prediction is known as *ab initio* gene discovery, which identify those common features using prediction software such as Genscan and N-Scan. The features might include the presence of long open reading frames in close proximity to transcriptional and translational initiation motifs and 3′ polyadenylation sites, and splicing consensus sequences (Gibson & Muse, 2009). We used the Genscan and N-Scan *ab initio* tools to predict genes from the epistatic QTL.

Genscan

Genscan is based on a general probabilistic model (generalized hidden Markov model) that predicts genes taking into consideration genomic information such as transcriptional, translational and splicing signals (Burge & Karlin, 1997). The features related to composition and length of introns, exons and intergenic regions are also considered in Genscan gene prediction. The algorithm scans both DNA strands for partial genes and complete genes. Genscan is one of the most widely used and most accurate *de novo* gene prediction tools with input of a single genome sequence in contrast to N-Scan, which requires at least two sequences for alignment and then prediction (Gross & Brent, 2006). This tool reveals only probable genes but provides no other information/annotation. Genscan is hosted at http://genes.mit.edu/GENSCAN.html. It can also be accessed from the UCSC Table Browser's Gene and Gene Prediction track in a more customized way.

N-Scan

N-Scan is also an *ab initio* gene prediction tool, which uses multiple alignment of sequences (obtained from an informant sequence aligned with one or more target sequences) along with the biological signals (common features of protein coding transcripts) in the target sequence (s) to predict genes (Gross & Brent, 2006). It creates a phylogenetic relationship from the multiple aligned sequences of target and informant sequences to predict genes. For chicken sequences, it uses Zebra Finch sequence as an informant sequence for alignment and hence gene prediction. It is provided by the Computational Genomics Lab at Washington University in St. Louis (http://mblab.wustl.edu/nscan/) for public use. Like Genscan it can also be accessed from the UCSC Table Browser in the same way.

After reviewing the number of genes and transcripts in the three epistatic QTL from all the databases and tools we used, we decided to select genes from the Ensembl and Genscan for further investigation into the identification of candidate genes because those revealed more genes and transcripts than RefSeq and N-Scan. We wanted to have as many genes as possible to start with the identification of candidate genes.

Combination of gene and conservation information

The information on gene prediction was combined with analysis of conservation patterns. For that purpose the most conserved regions were extracted in the epistatic QTL using the Comparative Genomics' Most Conserved track (phastConsElements7way) in the UCSC Table Browser. Using Galaxy's Text Manipulation and Operate on Genomic Intervals tools we identified those Ensembl and Genscan genes and transcripts that contained the most conserved genomic regions. The score used for most conserved elements was 500 or more (Score ranges 0-1000). We had two lists of those genes and transcripts considered conserved, one for Ensembl and one for Genscan.

The conserved Genscan and Ensembl genes/transcripts which are approximately the same in number share the overlapping genomic intervals as shown by the USCS Genome Browser (Figure 3, 4 and 5). Figure 2 exemplifies those overlaps between Genscan genes and Ensembl transcripts. Four genes in the same genome coordinates (1kbp-22kbp for example) are shown to have occupied the overlapping regions. Therefore, we decided to further investigate only Ensembl-annotated conserved genes and transcripts which have the advantage over Genscan for being well supported from previously annotated cDNA reference sequence or EST sequences. From here on we refer to these Ensembl conserved genes as "candidate genes". The Gene Ontology (GO) and KEGG pathways databases were analysed with these candidate Ensembl genes.

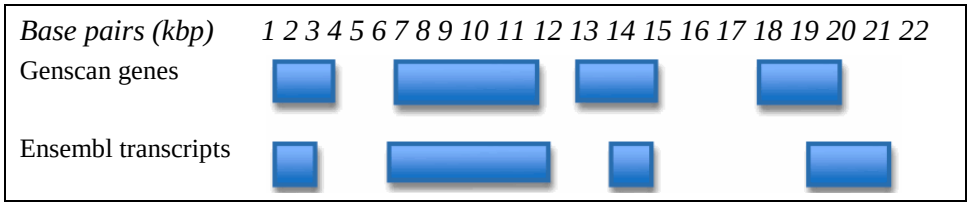


Figure 2. Genes explaining the Genscan and Ensembl overlap

Gene Ontology Consortium

The Gene Ontology (GO) Consortium (Ashburner et al, 2000) established a project to take care of the issue of inconsistent gene and gene product annotation terms among different gene product databases of different species. The same function used to have different term names, which made it difficult to search for researchers and even computer programmes for finding paralogy and orthology in different species. To remove that inconsistency of terms, the GO project introduced three basic gene ontology terms to describe gene products independent of the species. The genes and gene products now have their associated functions in three domains namely, Biological Processes (bp), Cellular Components (cc) and Molecular Functions (mf) in the GO database. The GO is freely accessible on the web at http://www.geneontology.org/. Main genome projects also provide the links to GO for specified genes. In Ensembl BioMart researchers can also extract gene ontology for the desired Ensembl genes. The GO project is constantly improving and maintaining its controlled ontology and annotation vocabulary for genes and their products for making it complete and accurate. The consortium has also developed set of tools, namely, Amigo (http://amigo.geneontology.org/) and OBO-Edit (http://oboedit.org/) for researchers to access GO database customized to their needs. Currently the GO database holds the gene annotations for a variety of species and for a large number of genes. One of the main applications of GO is the prediction of the association of a particular gene(s) with some diseases or traits of interest that are still unmapped to any genes. The GO consortium comprises many participating ontology databases. Some of those databases are species-specific and some are organ-specific while others are investigating specific biological systems. The curators use evidence codes to annotate the genes and gene products. Evidence codes are the terms explaining the support for their annotation of genes. Evidences are from experiments (inferred from experiment EXP), direct assays (inferred from direct assay IDA) and electronic annotation (inferred from electronic annotation IEA) to list a few. Most widely accepted GO is a unified bioinformatic initiative for constructing a structured vocabulary for species-neutral annotation of genes and gene products and the current genome projects are annotating genes by employing GO terms (Hennig et al, 2003; Raychaudhuri et al, 2002). Hennig et al. (2003) also concluded the correctness in most cases of annotation of sequence data using GO terms. The GO is the best in annotating the gene functions because the highly skilled biologists curate the genes after reviewing available information from literature and other such sources for evidence (Raychaudhuri et al, 2002). GO is arguably the most correct and popular ontology and annotation database to date.

Gene Ontology of the candidate genes

Using Ensembl BioMart the GO database was searched with the candidate Ensembl genes for all three kinds of gene annotation, namely, Biological Process, Molecular Function and Cellular Component. The annotations for the candidate Ensembl genes were studied along with KEGG pathways to understand the function of those genes in the biological pathways.

GenomeNet / KEGG

Kyoto University Bioinformatics Centre is hosting GenomeNet (http://www.genome.jp/), which was established in 1991 under the Human Genome Program of Japanese government and carries out research in the field of functional genomics and bioinformatics (Kanehisa et al, 2009). It provides database resources including KEGG Genes, KEGG Pathways, KEGG Disease and KEGG Brite to mention a few. The bioinformatic tools in the categories of Sequence Analysis, Genome Analysis and Chemical Analysis are also freely available for functional genomic research. Kyoto Encyclopaedia of Genes and Genomes (KEGG) is a part of GenomeNet database resources. We decided to use genomic information from KEGG Genes (http://www.genome.jp/kegg/genes.html) and system information from KEGG Pathways (http://www.genome.jp/kegg/pathway.html). The KEGG Pathway database contains biological pathway maps drawn manually mainly based on the molecular interaction and reaction networks information from published literature and it also contains maps for both normal and perturbed state (Kanehisa et al, 2009). Theses maps are broadly categorized into six global maps, which are Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases and Drug Development. These major categories are further subdivided into specific molecular systems. The manually drawn and curated pathways are based on supported research from the literature and are interactive. The process of KEGG Pathway Mapping is to map large-scale molecular datasets of genomics, transcriptomics, proteomics, and metabolomics to its pathway maps for interpreting biological functions. The maps are in the form of images in which mapped objects can be coloured differently from others. Most of the objects in the pathway maps are hyperlinked to detailed information from different KEGG databases. Links related to mapped objects are also provided to popular protein and genomic databases like Uniprot, Pfam, PROSITE, Ensembl and NCBI. Our purpose of using KEGG pathway maps was to map the available genomic information, candidate Ensembl genes in our case, to KEGG maps so that we could understand the role of those genes in those pathways and could be able to find any relationship of those pathways to complex quantitative trait of growth. Currently there are around 600 KEGG entries in Pubmed that shows its popularity among functional genomic researchers.

We tested some other biological pathway databases. One of those is REACTOME (http://www.reactome.org/) which is a knowledgebase of biological pathways and processes and provides information on molecular reaction systems (Vastrik et al, 2007). This web resource has less than 40 entries yet in Pubmed and do not provide the mapping support for the molecular datasets in contrast to KEGG.

We have selected a few pathways along with the candidate genes involved, in our hunt for the causative genes and mutations affecting phenotype of growth.

Summary of Materials and Methods

1. Bi-directionally selected chicken lines for body weight at 56days for more than 40 generations, the High-Line and the Low-Line (highly homozygous for QTL and genetic markers genotypes) are available for QTL analysis of growth

2. Genome re-sequencing data of High-Line and Low-Line for SNP detection is also available

3. The results from experimental crosses (F2 reciprocal Intercrosses) for QTL analysis of growth were:
   a. 13 QTL detected affecting growth with minor additive effects
   b. Epistasis between six QTL regions were found to play a major role in the regulation of growth

4. *Ab initio* discovery and database search in epistatic QTL regions for genes and transcripts using Ensembl, RefSeq, Genscan and N-Scan

5. Ensembl and Genscan genes selected over RefSeq and N-Scan as those contained more potential candidate genes

6. To identify conserved genes, the evolutionary most conserved sequences in chicken genome were determined in epistatic QTL

7. Identification of conserved Ensembl and Genscan genes containing the most conserved sequences in epistatic QTL regions

8. Conserved Ensembl genes were preferred over conserved Genscan genes because both gene tracks occupy approximately the same genomic intervals, are approximately the same in number and Ensembl genes are being better supported from gene evidence. Conserved Ensembl genes are now considered as candidate genes

9. Gene Ontology (GO) database consulted for functions of candidate genes from Ensembl

10. Candidate Ensembl genes were mapped to KEGG biological pathways.

RESULTS AND DISCUSSION

Genes

Our findings using Ensembl and RefSeq gene databases and Genscan and N-Scan gene prediction tools regarding the number of genes in the previously mapped interacting QTL for growth on chicken chromosome 3, 4 and 7 are given in Table 3.

Table 3. The size of QTL and the number of genes/transcripts therein

Chr / QTL	QTL (start bp-end bp)	Size of QTL (bp)	Ensembl Gene/Transcript	RefSeq Gene/Transcript	Genscan Gene	N-Scan Gene
3	28,768,446-39,739,740	10,971,295	140/182	34/142	265	198
4	9,401,881-15,000,000	5,598,120	118/141	26/112	194	124
7	16,446,768-28,614,633	12,167,866	214/316	63/208	419	297
Total			472/639	123/462	878	619

Table 3 shows Ensembl and RefSeq genes and transcripts (each gene having one or more transcripts). Ensembl contained more genes and transcripts than RefSeq mainly due to the feature of Ensembl gene predictions. The Ensembl database contains different sets of genes including known protein-coding, projected protein-coding, novel protein-coding, pseudogenes and RNA genes. Whereas the RefSeq is the collection of only known, well-annotated sequences of DNA, RNA and proteins, which are non-redundant and well supported from literature. Table 3 also shows that Genscan predicted more genes than N-Scan because of the different principles of gene discovery discussed already. Gene prediction tools like Genscan and N-Scan do not provide any evidence for the predicted genes and the prediction is only based on the preset parameters to look for probable genes in the given genomic sequences. In human there are 51,737 Ensembl genes in 3.27 Gbp of DNA (Ensembl database version 59.37d) whereas chickens have 17,934 Ensembl genes in 1.05 Gbp (Ensembl database version 59.2o). It is obvious that approximately three times more DNA in human than chicken also reveals approximately three times more genes in the Ensembl database. Approximately 63,300 bases represent one human gene while for chicken these bases are about 58,500. In our case 472 Ensembl genes were found in 28,737,281 bp of DNA representing approximately 60900 bases per gene.

The purpose of using Ensembl, RefSeq, Genscan and N-Scan for known genes and for gene predictions was to have as many genes as possible in three QTL before we actually started to exclude them and identify the candidate genes for growth.

Conservation pattern and conserved genes

In our study, the known and predicted Ensembl genes/transcripts and Genscan predicted genes were selected for the subsequent conservation analysis because those revealed more genes than RefSeq or N-scan. To identify the most likely candidate genes for growth we decided to select those genes that probably remained most conserved in evolution. The UCSC Genome Browser

and Table Browser provide the phastConsElements7way track for extracting the most conserved regions in the given QTL. The Most Conserved option in Comparative Genomics' track predicts conserved elements which are segments of alignments in QTL obtained from the pair-wise and then multiple alignment of 7 vertebrate species (chicken, human, mouse, rat, opossum, zebra fish and X. tropicalus) from 4 vertebrate classes namely, Fish, bird, mammals and amphibian. The Most conserved regions were scored having values from 0 to 1000. Table 4 shows the number of Most Conserved regions in all three QTL regions as 26,313 when the score was set equal to 0 and 3,118 when set equal to or greater than 500.

Both the Ensembl transcripts and Genscan genes were intersected separately with the most conserved regions (having conservation score ≥ 500) to identify those transcripts and genes that contain one or more of those conserved regions. The Galaxy's intersection tool was used to extract those genes. The Table 5 shows 459 Ensembl transcripts and 450 Genscan genes that contain at least one of the most conserved regions. From here on we refer to those genes as conserved genes.

Table 4: The number of most conserved regions in candidate QTL

Chr / QTL	Most Conserved Regions (Score ≥ 0)	Most Conserved Regions (Score ≥ 500)
3	7,787	864
4	4,535	482
7	13,991	1,772
Total	26,313	3,118

Table5: The number of genes/transcripts containing the most conserved regions (Score ≥ 500).

Chr / QTL	Ensembl Conserved Transcripts	Ensembl Conserved Genes	Genscan Conserved Genes
3	134	98	128
4	84	68	90
7	241	155	232
Total	459	321	450

Overlaps of genomic intervals in UCSC Genome Browser

When the conserved transcripts and genes tracks from Ensembl and Genscan were viewed in parallel in UCSC Genome Browser, most of the genes from both tracks appeared to occupy the overlapping genomic regions. Figures 1, 2 and 3 show those gene overlaps. This pattern suggested to primarily focusing on the Ensembl conserved genes as these have much more information and evidence for them to be probably the true genes (as described in the previous sections). To be raised to the gene status in genome annotation, the computationally predicted putative genes need to be confirmed by some supporting evidence from previously identified cDNA or EST sequences or known translated proteins. The Ensembl database is a collection of both known and putative genes that are well supported from second line of evidences. Therefore, Ensembl conserved genes were identified as candidate genes at that stage.

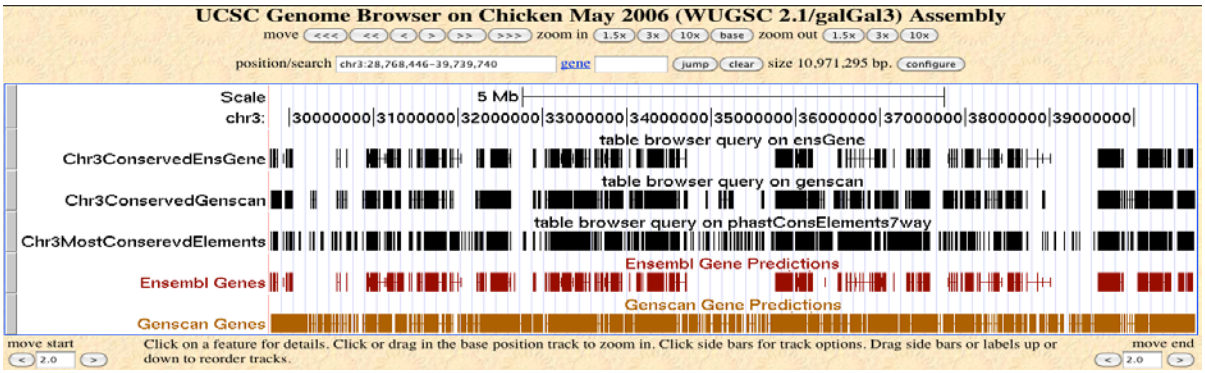


Figure 3: Chromosome 3 QTL Region showing genes and conservation in UCSC Genome Browser

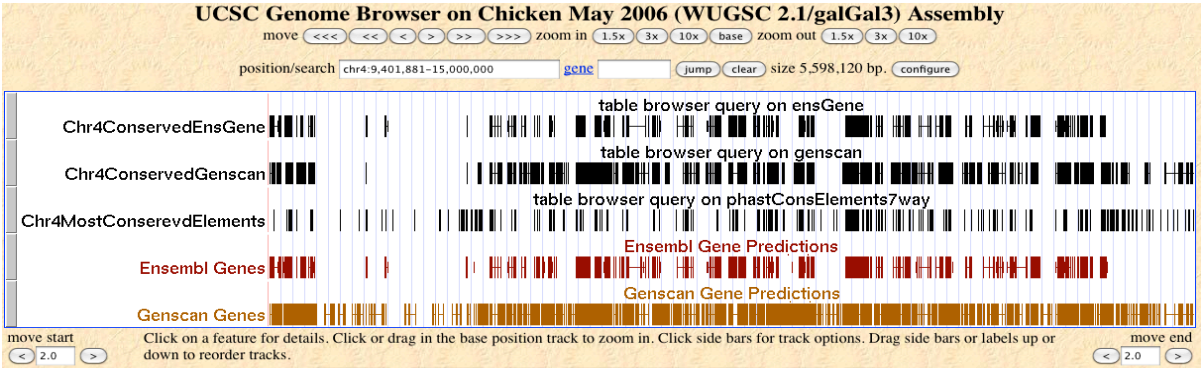


Figure 4: Chromosome 4 QTL Region showing genes and conservation in UCSC Genome Browser

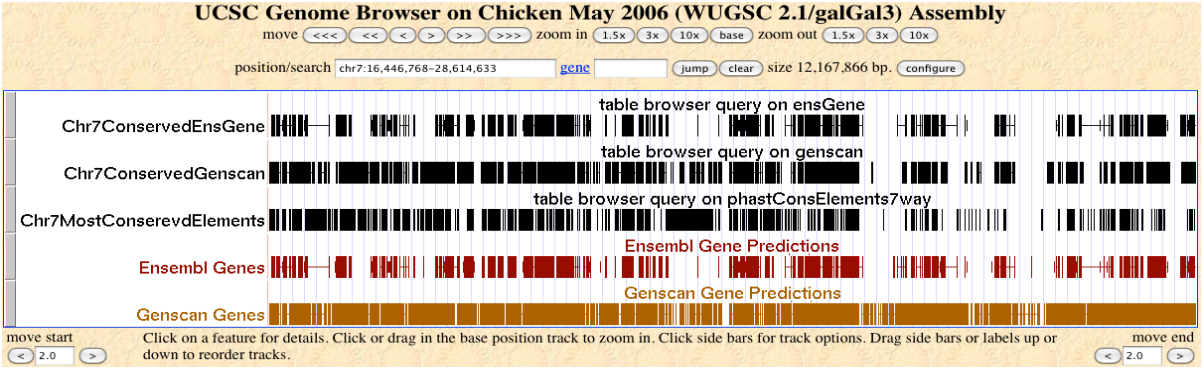


Figure 5: Chromosome 7 QTL Region showing genes and conservation in UCSC Genome Browser

Bioinformatic analysis of KEGG pathways and interactions

To further investigate the functions of candidate Ensembl genes we decided to map them to the KEGG biological pathway database. There are 147 chicken biological pathways in the KEGG Pathway database. We found many pathways in which candidate Ensembl genes were involved. Using Galaxy's tools like Text Manipulation, Filter and Sort, Join and Group we were able to identify those pathways that had gene representation from interacting QTL. A few of those pathways had the candidate genes from all three interacting QTL. Table 6 shows the actual number of Ensembl genes and the corresponding number of genes contained in KEGG gene database (KEGG gene). It also shows how many of the 147 biological pathways of chicken in the KEGG database were traced with those KEGG genes/Ensembl genes from all three candidate QTL. A total of 71 pathways were found with the candidate Ensembl genes and 76 pathways if we consider all Ensembl genes.

Table 6: Ensembl/KEGG genes from all three epistatic QTL and the related KEGG Biological Pathways in Chicken

Genes and Transcripts from three QTL	Ensembl Transcripts	Ensembl Genes	KEGG Genes	KEGG Genes mapped to KEGG Pathways	KEGG Pathways mapped with KEGG genes
Candidate	459	321	247	72	71
All	639	472	307	80	76

Table 7 shows the numbers of KEGG pathways, which may explain observed QTL interactions because they are represented with gene(s) in two or more QTL. A total of 24 and 27 such pathways were found when only candidate Ensembl genes were considered and when all Ensembl genes were considered, respectively.

Table 7: Number of KEGG pathways found which have genes represented in two or three of the epistatic QTL studied

Chr / QTL	KEGG Pathways (with candidate Ensemble genes)	KEGG Pathways (with all Ensembl genes)
3 and 7	6	5
3 and 4	7	8
4 and 7	5	6
3, 4 and 7	6	8
Total	24	27

The KEGG biological pathways involving our candidate Ensembl genes present a way forward in our investigation of genetic architecture underlying phenotypes of our interest. We have found 6 biological pathways having genes from all of the interacting QTL on chromosomes 3, 4 and 7. Table 8 briefly explains the major functions of those 6 biological pathways. Candidate Ensembl genes from those QTL involved in those 6 pathways are 35 in number. Those 35 candidate genes are also involved in 40 other pathways as one gene may be involved in more than one pathway. Table 9 shows those 6 pathways along with the candidate genes.

Here we discuss the Mitogen Activated Protein Kinase (MAPK) signalling pathway, which involves candidate Ensembl genes from all of the three epistatic QTL. The MAPK cascade includes a series of highly conserved protein kinases, which are considered to be involved in regulation of cell proliferation in eukaryotes. Once these kinases are inside the nucleus these also affect homeostasis, cell growth and viability/apoptosis by altering gene expression. The signalling of MAPK pathway results from three successive phosphorylation events, called MAPK cascade. MAPKs phosphorylations result in distinct functions for its substrates, which also include mitogen activated proteins (Faustino et al. 2010; Shiryaev & Moens, 2010; Huang, 1996). The MAPK involves seven of the Ensembl candidate genes from candidate QTL listed in Table 7.

Gene annotations from GO for the candidate genes involved in the MAPK signalling pathway show that most genes are involved in metabolism related to phosphorylation of the kinases. That phosphorylation metabolism is a hallmark of MAPK signalling cascade that affects cellular growth, proliferation, differentiation and development. One of the genes (Q197G3_CHICK/ ENSGALG00000007806) has a growth factor activity related to it and another gene (ATF2_CHICK/ENSGALG00000009287) is involved in regulation of transcription in general, which affects gene expressions and might also affect growth eventually.

The Adipocytokine signalling pathway is also an interesting pathway as it has functional roles in regulating appetite. The Low-Line chickens have low appetite after generations of selection. Therefore, an investigation into this pathway along with the candidate genes it involves may reveal the genetic basis of low appetite in Low-Line. This pathway controls the production of leptin and adiponectin hormones through the volume and the number of adipocytes. Adipocytes volume positively correlates the leptin production. Leptin, by acting at hypothalamic nuclei and affecting the levels of neuropeptides (like NPY, AGRP, and alpha-MSH) regulates metabolic rate and energy intake. Leptin in adipose tissue regulates energy intake through feeding and expenditure of energy in the body. Appetite suppressing actions of leptin are hampered in obesity (Munzberg & Myers, 2005). Obesity has been identified to associate with resistance to leptin (Becskei et al, 2010).

We are also investigating other pathways having candidate genes from epistatic QTL. Currently MAPK and Adipocytokine signalling pathways are the key candidates and are being investigated.

Table 8: Major functions of KEGG pathways having gene representation from all three candidate QTL

Pathway ID	Pathway Name	Major Functions
gga01100	Metabolic Pathways	Metabolic pathways include sub-pathways that are involved in biosynthesis and degradation of all sorts of biological substances. The changes in the rate of metabolism may affect the growth rate in general.
gga04010	MAPK Signalling Pathway	The MAPK signalling pathway is involved in numerous cellular functions which includes cell proliferation, cell differentiation and cell growth.
gga04510	Focal Adhesion	Focal adhesions are formed at contact points of cell-extracellular matrix and have different important biological roles such as cell proliferation, cell differentiation and survival of the cell. They are also involved in regulation of gene expression.
gga04530	Tight Junction	Tight junction is involved in cellular processes such as cell communication. The epithelial tight junction proteins form diffusion barriers at intramembrane and paracellular level.
gga04810	Regulation of Actin Cytoskeleton	The actin cytoskeleton helps regulate the shape of the cell and its motility, which is important to the development of biological systems such as central nervous system.
gga04920	Adipocytokine Signalling Pathway	The Adipocytokine signalling pathway controls the production of leptin and adiponectin through the volume and number of adipocytes. Leptin imparts anorectic condition by affecting the levels of neuropeptides.

Table 9: Pathways and Ensembl candidate genes involved

Pathway ID	Pathway Name	Chr / QTL	Ensembl Gene ID	KEGG Gene ID
gga01100	Metabolic pathways (16)	3	ENSGALG00000010349	gga:421452
		3	ENSGALG00000010557	gga:421467
		3	ENSGALG00000010702	gga:428579
		3	ENSGALG00000010768	gga:428583
		3	ENSGALG00000014464	gga:422069
		4	ENSGALG00000007493	gga:422302
		4	ENSGALG00000007793	gga:422327
		4	ENSGALG00000007936	gga:395833
		4	ENSGALG00000007990	gga:422339
		4	ENSGALG00000008088	gga:422345
		7	ENSGALG00000009236	gga:424135
		7	ENSGALG00000009286	gga:424142
		7	ENSGALG00000009589	gga:395743
		7	ENSGALG00000010931	gga:429027
		7	ENSGALG00000010960	gga:424177
		7	ENSGALG00000011246	gga:769112
gga04010	MAPK signalling pathway (7)	3	ENSGALG00000010435	gga:421460
		3	ENSGALG00000010709	gga:421497
		4	ENSGALG00000007806	gga:422330
		7	ENSGALG00000009287	gga:395727
		7	ENSGALG00000009344	gga:424149
		7	ENSGALG00000011502	gga:424225
		7	ENSGALG00000011592	gga:395149
gga04510	Focal adhesion (8)	3	ENSGALG00000010290	gga:395909
		3	ENSGALG00000010709	gga:421497
		3	ENSGALG00000014463	gga:396263
		4	ENSGALG00000007646	gga:422316
		4	ENSGALG00000008058	gga:422342
		4	ENSGALG00000008266	gga:422350
		7	ENSGALG00000009362	gga:396226
		7	ENSGALG00000011141	gga:424191
gga04530	Tight junction (6)	3	ENSGALG00000010385	gga:421455
		3	ENSGALG00000010709	gga:421497
		3	ENSGALG00000014463	gga:396263
		4	ENSGALG00000007311	gga:422292
		7	ENSGALG00000009344	gga:424149
		7	ENSGALG00000011592	gga:395149
gga04810	Regulation of actin cytoskeleton (8)	3	ENSGALG00000010778	gga:396364
		3	ENSGALG00000014463	gga:396263
		4	ENSGALG00000007806	gga:422330
		4	ENSGALG00000008058	gga:422342
		7	ENSGALG00000009362	gga:396226
		7	ENSGALG00000011141	gga:424191
		7	ENSGALG00000011452	gga:429041
		7	ENSGALG00000011592	gga:395149
gga04920	Adipocytokine signalling pathway (3)	3	ENSGALG00000010709	gga:421497
		4	ENSGALG00000008088	gga:422345
		7	ENSGALG00000011360	gga:424208

The MAPK Signalling Pathway

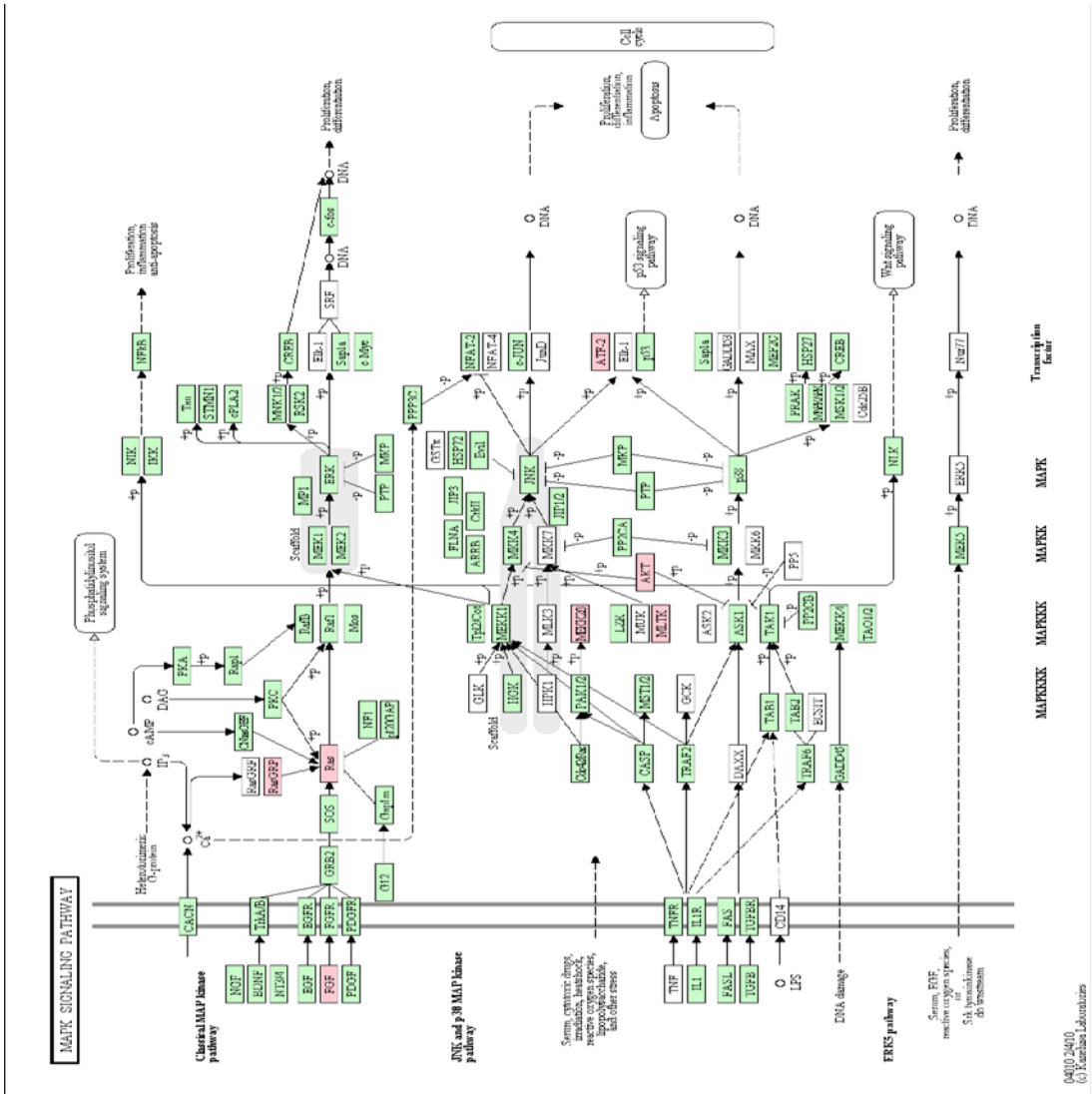


Figure 6. The MAPK signalling pathway in KEGG database, showing the candidate genes and gene products (coloured pink). (Reproduced with permission)

The Adipocytokine signalling pathway

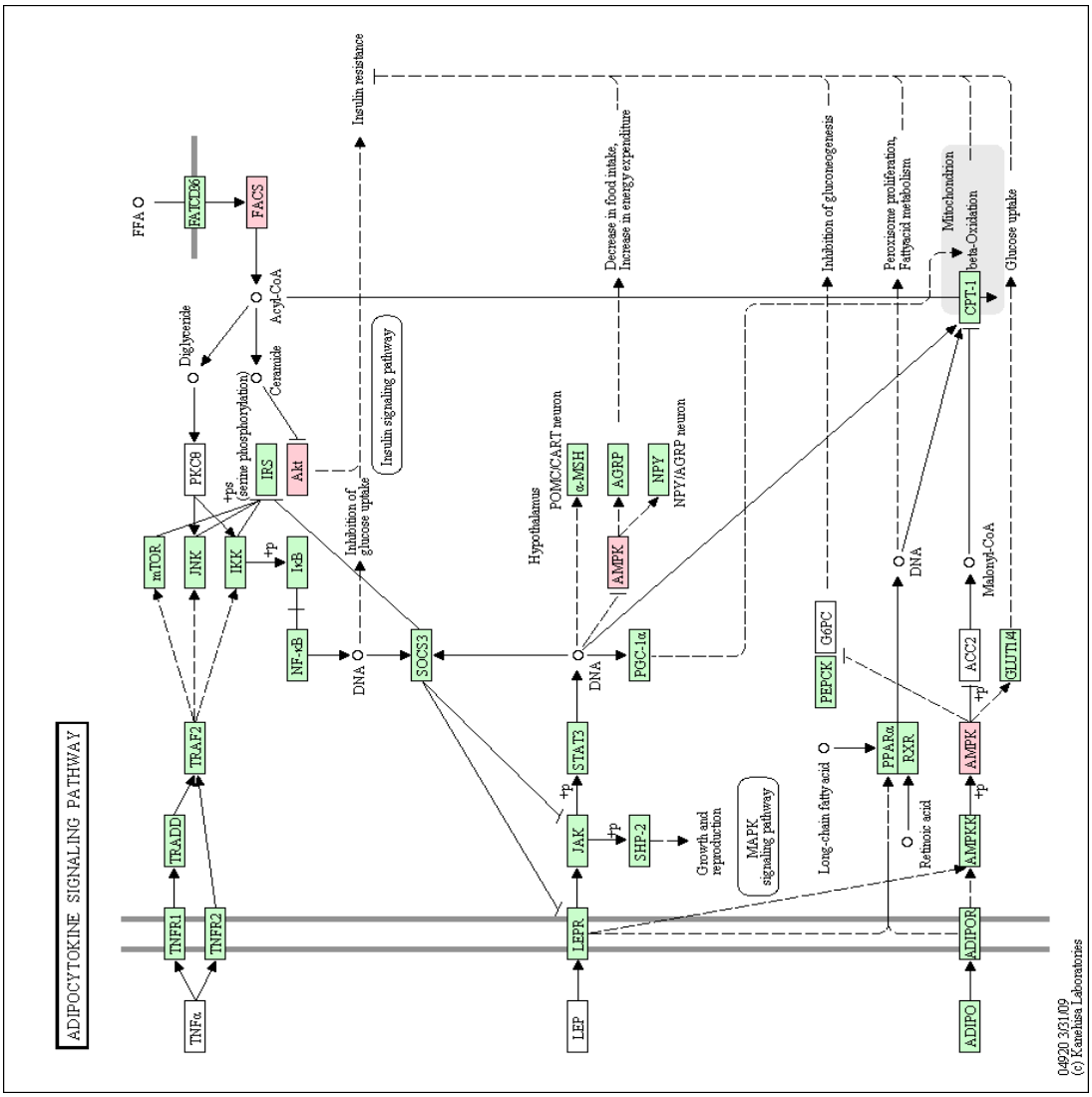


Figure 7. The Adipocytokine signalling pathway in KEGG database, showing the candidate genes and gene products (coloured pink). (Reproduced with permission)

Searching for the causative mutation(s)

Polymorphism in the candidate genes of both chicken lines will be functionally evaluated, potentially explaining the genetic mechanism underlying the phenotypic differences between those lines. Table 9 shows the number of SNPs detected in genome re-sequencing in different parts of Ensembl genes in the epistatic QTL.

Table 9: SNP distribution across different parts of all Ensembl genes in candidate QTL

QTL→ Gene Region↓	Chr 3	Chr 4	Chr 7	Total SNPs	SNPs (per kb)
Genes	23,249	9,811	32,736	65,796	3.66
CpG island	229	237	523	989	2.42
Exons	711	552	1,259	2,522	2.07
Introns	22,538	9,259	31,484	63,281	3.78
5’ UTR	21	25	41	87	2.40
3’ UTR	154	173	180	507	2.85
Coding exons	536	354	1,038	1,928	1.91
5bp (exons)	43	33	62	138	2.01
5bp (introns)	36	16	79	131	2.10

The SNP distribution shows that most SNPs are in intronic regions whereas the exons contain a very low percentage of those SNPs. Important mutations, which contribute to the divergent phenotypes in the High- and Low Chicken Lines might be residing in coding sequences and altering protein structure and function or might be located in regulatory regions and affect gene expression or post-translational processes.

FUTURE PROSPECTS

The specific aim of the current project was to explore the genome for its genes in the previously mapped QTL for growth in general in chicken. A recent fine mapping study using an Advanced Intercross Line (AIL) confirmed QTL on chromosome 1, 2, 3, 4, 5, 7 and 20 controlling growth (Besnier et al, submitted). We have found a few candidate genes in the three important interacting QTL on chromosome 3, 4 and 7. In a near future we would scan the remaining detected QTL on chromosome 1, 2, 5, and 20 for the conserved genes and map them to interaction pathways, hopefully finding new candidate genes.

After identifying the real candidate genes the polymorphism therein would be investigated. These investigations require efficient techniques to develop bioinformatic and computational algorithms, methods and tools to reveal and extract the interesting information from biological databases available for public use. Thus, the computational development done primarily for this project will be of broader interest and are likely to benefit many other research projects.

To identify protein-coding sequences, regulatory regions, conserved sequences a pipeline would be developed using bioinformatic resources and scripting languages. Phylogenetic analysis needs to be done to identify orthologs and hence shared functional information with the chicken genes.

In the future we shall use genome re-sequencing and 60K SNP chip genotyping data for computational analysis to identify selective sweeps within the detected QTL.

The identification of the polymorphism in functional elements with prediction and ranking of the possible phenotypic contribution for mutations and interaction patterns may reveal and explain genetic mechanism underlying the phenotypic differences between the High- and Low Chicken Lines. Such findings could significantly contribute to our understanding of metabolic pathways regulating growth, which may benefit animal breeding, human medicine and/or other areas of biology.

Finally the identified candidate mutations and pathways would be tested experimentally for their presumed functional effect on phenotypic variation between the High- and the Low chicken lines.

ACKNOWLEDGEMENTS

In the name of Allah, Most Gracious, Most Merciful. Thanks to Almighty Allah.

This research work was carried out in Computational Genetics group at the Department of Animal Breeding and Genetics at SLU. I really want to thank all the members of the group for extending all kind of support during this work. Special thanks to Stefan Marklund for supervising this work and correcting all my mistakes with so much patience. Also thanks to Marcin and Xidan for their valuable suggestions. I anticipate the same gestures from them in the future.

I would like to thank Higher Education Commission of Pakistan (HEC) for funding this research work through a scholarship programme, which is aimed at promoting advancement of universities and institutes of higher learning in Pakistan. The aim of the program is to strengthen Pakistan's national research capacity through educational cooperation between Pakistan and Sweden.

At this moment I really want to thank Erik Bongcam-Rudloff for all his support and coordination with HEC and Swedish Institute (SI). He always came up with the solutions to all kinds of problems we faced from time to time. We are looking forward to the same kind of support in the coming years.

My parents in Pakistan have been praying humbly all the time to The Almighty ALLAH for the success of all their children in all fields of life. I really love them.

I want to thank my family for their support. I love my family and I can't imagine a moment without them. I love my wife Madeeha for her love, care and all her support. I love my two little cute daughters Ayeza and Fiza too. I want them around all the time. They are really wonderful and the joy of my life.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet. May;25(1):25-9

Barnes MR. 2010. Exploring the landscape of the genome. Methods Mol Biol. 628:21-38.

Becskei C, Lutz TA, Riediger T. 2010. Reduced fasting-induced activation of hypothalamic arcuate neurons is associated with hyperleptinemia and increased leptin sensitivity in obese mice. Am J Physiol Regul Integr Comp Physiol. Jun 10

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, .L, Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., Miller, W. 2004. Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner. Genome Res. 14(4), 708-15.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. Jan; Chapter 19:Unit 19.10.1-21.

Broman KW. 2001. Review of statistical methods for QTL mapping in experimental crosses. Lab Anim (NY). Jul-Aug;30(7):44-52

Burge, C. and Karlin, S. 1997.Prediction of complete gene structures in human genomic DNA. 1997. J. Mol. Biol. Apr 25;268(1):78-94

Carlborg Ö, Jacobsson L, Ahgren P, Siegel P, Andersson L. 2006. Epistasis and the release of genetic variation during long-term selection. Nat Genet. Apr;38(4):418-20.

Carlborg Ö. 2002. New Methods for Mapping Quantitative Trait Loci. Doctoral dissertation. ISSN 1401-6257, ISBN 91-576-6362-9

Cheema MA, Qureshi MA, Havenstein GB. 2003. A comparison of the immune response of a 2001 commercial broiler with a 1957 randombred broiler strain when fed representative 1957 and 2001 broiler diets. Poult Sci.Oct;82(10):1519-29.

Chiaromonte, F., Yap, V.B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. Pac Symp Biocomput 2002, 115-26.

Dunnington EA, Siegel PB. 1996. Long-term divergent selection for eight-week body weight in white Plymouth rock chickens. Poul Sci. Oct;75(10): 168-79.

Faustino RS, Maddaford TG, Pierce GN. 2010. Mitogen activated protein kinase at the nuclear pore complex. J Cell Mol Med. May 24Felsenstein, J. and Churchill, G.A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13, 93-104.

Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A,

Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Smith J, Searle SM. 2010. Ensembl's 10th year. Nucleic Acids Res. Jan;38(Database issue):D557-62.

Glazier AM, Nadeau JH, Aitman TJ. 2002. Finding genes that underlie complex traits. Science. Dec 20;298(5602):2345-9.

Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. Aug 25;11(8):R86.

Greg Gibson; Spencer V. Muse. 2009. A Primer of Genome Science, Third Edition. Sinauer Associates; Pages: 344

Gross SS, Brent MR. 2006. Using multiple alignments to improve gene prediction. J Comput Biol. Mar;13(2):379-93.

Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A.2009. BioMart Central Portal--unified access to biological data. Nucleic Acids Res. Jul 1;37(Web Server issue):W23-7.

Haley CS, Knott SA, Elsen JM. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics. Mar;136(3):1195-207.

Hannah Pulker, María Victoria Lareu, Christopher Phillips and Angel Carracedo. 2007. Finding genes that underlie physical traits of forensic interest using genetic tools. Forensic Sci Int Genet. Jun;1(2):100-4.

Hennig S, Groth D, Lehrach H. 2003. Automated Gene Ontology annotation for anonymous sequence data. Nucleic Acids Res. Jul 1;31(13):3712-5.

Huang CY, Ferrell JE. 1996. Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci U S A Sep;93(19):10078-83.

Jacobsson L, Park HB, Wahlberg P, Fredriksson R, Perez-Enciso M, Siegel PB, Andersson L. 2005. Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. Genet Res. Oct;86(2):115-25.

Jacobsson L, Park HB, Wahlberg P, Jiang S, Siegel PB, Andersson L. 2004. Assignment of fourteen microsatellite markers to the chicken linkage map. Poult Sci. Nov;83(11):1825-31.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. Jan;38(Database issue):D355-60.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. Jan 1;34(Database issue):D354-7.

Kanehisa, M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. Jan 1;28(1):27-30.
Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. Jan 1;32 (Database issue):D493-6

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. 2002. UCSC Genome Browser. Genome Res. 2002 Jun;12(6):996-1006.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100(20), 11484-11489.

Lander ES, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. Jan;121(1):185-99.

Liu P, Vikis H, Lu Y, Wang D, You M. 2007. Large-scale in silico mapping of complex quantitative traits in inbred mice. PLoS One. Jul 25; 2(7):e651.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep 15;437(7057):376-80.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet. Jan;11(1):31-46.

Morozova O, Marra MA. 2008. Application of next-generation sequence technologies in functional genomics. Genomics. Nov;92(5):255-64.

Murphy, W.J., et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294(5550), 2348-51.

Münzberg H, Myers MG Jr. 2005. Molecular and anatomical determinants of central leptin resistance. Nat Neurosci. May;8(5):566-70.

Parker, Glenn Proctor, James Smith, and Stephen M. J. Searle. 2010. Ensembl's 10th year. Nucleic Acids Research 38 Database issue:D557-D562

Paul Flicek, Bronwen L. Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karyn Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosée M. Fernández-Suarez, Javier Herrero ,Tim J. P. Hubbard, Anne

Pengyuan Liu, Haris Vikis, Yan Lu, Daolong Wang, Ming You . 2007. Large-Scale In Silico Mapping of Complex Quantitative Traits in Inbred Mice. PLoS ONE 2(7): e651. doi:10.1371/journal.pone.0000651.

Pevsner J. 2009. Analysis of genomic DNA with the UCSC genome browser. Methods Mol Biol. 537:277-301.

Randolph S. Faustino, Thane G. Maddaford, and Grant N. Pierce. 2010. Mitogen activated protein kinase at the nuclear pore complex. Journal of Cellular and Molecular Medicine

Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. 2002. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res. Jan;12(1):203-14.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. Jan; 38(Database issue):D613-9.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. Mar 25;464(7288):587-91.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W. 2003. Human-Mouse Alignments with BLASTZ. Genome Res. 13(1), 103-7.

Seger R, Krebs EG. 1995. The MAPK signaling cascade. FASEB J. Jun;9(9):726-35.11.

Shiryaev A, Moens U. 2010. Mitogen-activated protein kinase p38 and MK2, MK3 and MK5: ménage à trois or ménage à quatre? Cell Signal. Aug;22(8):1185-92.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionary conserved elements in vertebrates, insect, worm, and yeast genomes. Genome Res. Aug;15(8):1034-50.

Siepel, A. and Haussler, D. 2005. Phylogenetic hidden Markov models. In R. Nielsen, ed., Statistical Methods in Molecular Evolution, pp. 325-351, Springer, New York.

Taylor J, Schenck I, Blankenberg D, Nekrutenko A. 2007. Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics. Sep;Chapter 10:Unit 10.5.

Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. 2007. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 8(3): R39.

Woollard PM. 2010. Asking complex questions of the genome without programming. Methods Mol Biol. 628:39-52.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139, 993-1005.